



# Large-Scale Quality Analysis of Published ChIP-seq Data

## Citation

Marinov, Georgi K., Anshul Kundaje, Peter J. Park, and Barbara J. Wold. 2014. "Large-Scale Quality Analysis of Published ChIP-seq Data." *G3: Genes|Genomes|Genetics* 4 (2): 209-223. doi:10.1534/g3.113.008680. <http://dx.doi.org/10.1534/g3.113.008680>.

## Published Version

doi:10.1534/g3.113.008680

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:11879885>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

# Large-Scale Quality Analysis of Published ChIP-seq Data

Georgi K. Marinov,<sup>\*</sup> Anshul Kundaje,<sup>\*\*,††,1</sup> Peter J. Park,<sup>†,\*,§</sup> and Barbara J. Wold<sup>\*,2</sup>

<sup>\*</sup>Division of Biology, California Institute of Technology, Pasadena, California 91125, <sup>†</sup>Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, <sup>‡</sup>Informatics Program, Children's Hospital Boston, Boston, Massachusetts 02115, <sup>§</sup>Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, <sup>\*\*</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and <sup>††</sup>The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142

**ABSTRACT** ChIP-seq has become the primary method for identifying *in vivo* protein–DNA interactions on a genome-wide scale, with nearly 800 publications involving the technique appearing in PubMed as of December 2012. Individually and in aggregate, these data are an important and information-rich resource. However, uncertainties about data quality confound their use by the wider research community. Recently, the Encyclopedia of DNA Elements (ENCODE) project developed and applied metrics to objectively measure ChIP-seq data quality. The ENCODE quality analysis was useful for flagging datasets for closer inspection, eliminating or replacing poor data, and for driving changes in experimental pipelines. There had been no similarly systematic quality analysis of the large and disparate body of published ChIP-seq profiles. Here, we report a uniform analysis of vertebrate transcription factor ChIP-seq datasets in the Gene Expression Omnibus (GEO) repository as of April 1, 2012. The majority (55%) of datasets scored as being highly successful, but a substantial minority (20%) were of apparently poor quality, and another ~25% were of intermediate quality. We discuss how different uses of ChIP-seq data are affected by specific aspects of data quality, and we highlight exceptional instances for which the metric values should not be taken at face value. Unexpectedly, we discovered that a significant subset of control datasets (*i.e.*, no immunoprecipitation and mock immunoprecipitation samples) display an enrichment structure similar to successful ChIP-seq data. This can, in turn, affect peak calling and data interpretation. Published datasets identified here as high-quality comprise a large group that users can draw on for large-scale integrated analysis. In the future, ChIP-seq quality assessment similar to that used here could guide experimentalists at early stages in a study, provide useful input in the publication process, and be used to stratify ChIP-seq data for different community-wide uses.

## KEYWORDS

ChIP-seq  
chromatin  
immuno-  
precipitation  
cross-correlation  
quality  
assessment  
transcription  
factor

Copyright © 2014 Marinov *et al.*

doi: 10.1534/g3.113.008680

Manuscript received September 29, 2013; accepted for publication November 21, 2013; published Early Online December 17, 2013.

This is an open-access article distributed under the terms of the Creative Commons Attribution Unported License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supporting information is available online at <http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.113.008680/-/DC1>

A global analysis of published ChIP-seq data addressing a subset of the quality control issues presented here was published while this paper was under final review (Mendoza-Parra *et al.* 2013).

<sup>1</sup>Present address: Department of Genetics, Department of Computer Science, Stanford University, Stanford, CA 94305.

<sup>2</sup>Corresponding author: Barbara J. Wold. Division of Biology, California Institute of Technology, 1200 E. California Boulevard, MC 156-29, Pasadena, CA 91125.

Chromatin immunoprecipitation (ChIP) (Gilmour and Lis 1984; Gilmour and Lis 1985; Solomon *et al.* 1988) experiments identify sites of occupancy by specific transcription factors (TFs), cofactors, and other chromatin-associated proteins as well as histone modifications. Such proteins are concentrated at specific loci via direct binding to DNA or by indirect binding mediated by other proteins or RNA molecules. In most ChIP protocols, proteins are first cross-linked to DNA, most often using formaldehyde. The fixed chromatin is sheared, and an antibody specific for the protein or histone modification of interest is used to retrieve protein:DNA complexes from which the DNA segments are released and then assayed. The assay was first applied to individual TF/promoter complexes by using qPCR to detect enrichment over specific DNA segments (Hecht *et al.* 1996). Subsequent adaptations extended it to large sets of promoters or other genomic regions by using microarrays (ChIP-on-Chip/ChIP-Chip) (Ren *et al.*

2000; Iyer *et al.* 2001; Lieb *et al.* 2001; Horak and Snyder 2002; Weinmann *et al.* 2002). Ultimately, the entire genome became accessible with the advent of high-throughput sequencing and the development of ChIP-seq (Johnson *et al.* 2007; Barski *et al.* 2007; Mikkelsen *et al.* 2007; Robertson *et al.* 2007).

In all cases, preferential enrichment of a given immunoprecipitated DNA segment is detected and quantified by comparing it with a control experiment in which there is no specific antibody enrichment step. These controls can be generated from sonicated DNA before immunoprecipitation (input) or a mock immunoprecipitation with an unrelated antibody (IgG). Sequencing-based ChIP has become the method of choice because it enables genome-wide coverage, even for large genomes, and because of its superior signal-to-noise characteristics compared to alternative methods. Since its initial development, ChIP-seq has been used in hundreds of publications (778 in PubMed as of December 18, 2012), including by the ENCODE consortium (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012), to map occupancy over 100 human TFs and cofactors in a diverse collection of cell lines (Gerstein *et al.* 2012; Wang *et al.* 2012).

A basic question for any ChIP-seq experiment is, how successful was it? It has taken several years for the field to develop objective ways to quantify key aspects of success in immunoprecipitation enrichment, library building, and final sequencing. Poor datasets that have high false-negative rates in peak calling are a predictable pitfall that has significant downstream consequences for some kinds of biological and computational analyses. For example, when lower-quality datasets are used for integrative analyses that are sensitive to false-negative rates, incorrect inferences and conclusions become likely (see *Discussion*). In estimating data quality, the traditional approach of visual inspection at a limited number of sites (often previously well-characterized using low-throughput approaches) is inefficient, subjective, and ultimately can be deceptive. It is also possible (and commonly observed in practice) that sites, the biological importance of which has been defined by independent functional assays, can decrease to below the sensitivity threshold of a poor or mediocre ChIP-seq experiment. Moreover, there is no current way to predict, *a priori*, the number of sites in the genome that should be detectable for a given factor and cell type. Most TFs studied thus far reproducibly occupy thousands to tens of thousands of sites (ENCODE Project Consortium 2012; Landt *et al.* 2012). Thus, a dataset for which several thousand sites have been called might in fact be capturing a minority of true positive interactions, or it might encompass virtually all biologically pertinent sites. To help address the problem of data assessment as part of the ENCODE project, we and others developed a set of ChIP-seq quality control (QC) metrics and guidelines (Landt *et al.* 2012) that were adopted and applied to all of its datasets. Substandard datasets were consequently replaced, flagged as substandard, and/or removed from analysis (ENCODE Project Consortium 2012; Landt *et al.* 2012).

Incorporating published datasets into an ongoing study can bring new biological insights and avoid unnecessary duplication of work. Variable quality of published data can be a significant barrier to these uses of existing data. They are the products of work from many different laboratories with invaluable expertise in specific biological systems, but they also use many variations of ChIP-seq experimental protocols and bioinformatics treatments. The extent and nature of the variations have not been assessed globally and systematically. In this work, we examined the GEO submission series containing vertebrate TF ChIP-seq datasets and found that ~20% of datasets scored as being of low quality, with an additional ~25% exhibiting intermediate ChIP enrichment. We also noticed that approximately one-third of studies have control datasets with a high degree of read clustering that

is normally expected only in ChIP-seq datasets. This was observed more often for the IgG control design than for input DNA controls. These and related observations argue for data quality measures routine characterization and reporting of ChIP-seq data quality measures.

## MATERIALS AND METHODS

### Sequencing read alignment

Raw sequencing reads for all non-ENCODE GEO series containing ChIP-seq datasets against TFs and chromatin-modifying proteins (submitted before April 1, 2012) were downloaded from GEO in SRA format and converted to FASTQ format using the fastq-dump program in the sratoolkit (version 2.1.9). Reads were aligned using Bowtie (Langmead *et al.* 2009) version 0.12.7 with the following setting: “-v 2 -t -k 2 -m 1 -best-strata,” which allows for two mismatches relative to the reference and only retains unique alignments. Human datasets were mapped against the male set of chromosomes (excluding all random chromosomes and haplotypes) for version hg19 of the human genome; the mm9 version of the mouse genome was used for mouse data, rn5 was used for rat data, danRer7 was used for zebrafish data, susScr2 was used for pig data, and xenTro3 was used for the clawed frog *Xaenopus tropicalis* data, and all assemblies were downloaded from the UCSC genome browser (Kent *et al.* 2002).

### ChIP quality assessment

ChIP quality assessment was performed on both ChIP and input datasets using the general strategy described by Landt *et al.* (2012). Because a library may score as an “unsuccessful ChIP” for reasons other than IP failure (e.g. being performed in a knockout background, in si/shRNA-treated cells, or in conditions under which the factor is not expressed or not bound to DNA), the following additional criteria were used to determine whether each library is expected to score positively in the QC assessment:

1. All experiments claimed to be successful by authors are expected to exhibit high level of read clustering.
2. All inputs (sonicated DNA and IgG mock IPs) are expected to exhibit minimal read clustering (QC tag of -2 or -1).
3. All ChIP-seq experiments performed in a knockout background for the factor are expected to exhibit minimal read clustering (QC tag of -2 or -1).
4. Because knockdown efficiency varies and because it is unknown what protein levels would be sufficiently high for the factor to be successfully ChIP-ed, ChIP-seq experiments performed in cells treated with si/shRNAs targeting the factor are set aside as “unknown” and assessed for library complexity and sequencing depth but not for ChIP quality.
5. Experiments against factors known to bind to DNA on some stimulus performed in unstimulated cells are also tagged as “unknown” because lower-level binding in unstimulated cells cannot be ruled out (and is, in fact, often observed).
6. Experiments performed in conditions that may result in the factor not binding to DNA (time courses, knockdowns, or knockouts for other factors that may affect binding of the targeted factor) are also tagged as “unknown.”
7. Other experiments not matching any of these categories are expected to exhibit high levels of read clustering.

Cross-correlation analysis was performed using version 1.10.1 of SPP (Kharchenko *et al.* 2008) and the following parameter: “-s = 0:2:400.” QC scores were assigned based on the relative strand

correlation (RSC) values (integers ranging from  $-2$  to  $2$ ,  $RSC \in \{0, 0.25\} \Rightarrow QC \leftarrow -2$ ,  $RSC \in \{0.25, 0.50\} \Rightarrow QC \leftarrow -1$ ,  $RSC \in \{0.50, 1.00\} \Rightarrow QC \leftarrow 0$ ,  $RSC \in \{1, 1.50\} \Rightarrow QC \leftarrow +1$ ,  $RSC \geq 1.5 \Rightarrow QC \leftarrow +2$ , with  $-2$  corresponding to minimal read clustering and  $2$  corresponding to a highly clustered library) and used as a measure of ChIP quality. These scores capture the extent of read clustering in a ChIP-seq experiment in organisms whose genomes have similar size and structure to those of mammals. We point out that these scores may not be appropriate in genomes with very different size and/or structure. This motivated us to discard data from nonvertebrate model organisms for this analysis. Different values than those used here for RSC or normalized strand correlation (NSC) coefficients may be needed for such genomes, and this is a topic for future investigation. Cross-correlation plots were manually examined to ensure no artifactual QC scores were included because of size selection issues (such as, for example, a library being fragmented to an average size close to the read length and confusing the automated fragment peak assignment). In general, we recommend manual examination of cross-correlation plots in all cases. This presents a deeper and more detailed view of the characteristics of the dataset because the cross-correlation profile provides not only information regarding ChIP enrichment but also regarding the fragment length distribution in the datasets. For example, a dataset might exhibit periodicity in the distribution of fragment size lengths, presenting itself as numerous smaller peaks along the curve (often seen when chromatin is enzymatically digested rather than sonicated), or it can deviate from the standard unimodal pattern (aside from the phantom peak) indicating issues with size selection. The code for running SPP and assigning QC scores is available at <https://code.google.com/p/phantompeakqualtools/>.

### MyoD and myogenin ChIP-seq peak calling

MyoD and myogenin datasets were generated by the Wold laboratory and are available under GEO accession number GSE44824. We note that the apparent weakness of the “myogenin 2” ChIP dataset is most likely attributable to undersequencing and would be elevated to high-quality status if sequenced deeper; undersequencing is one possible reason for suboptimal quality metrics (A. Kundaje *et al.*, unpublished data). Reads were mapped as described above and peaks were called using ERANGE3.2 (Johnson *et al.* 2007) with the following settings: “ $-\text{minimum } 2 -\text{ratio } 3 -\text{shift learn } -\text{revbackground } -\text{listPeak}.$ ” ChIP-seq peak calls were counted as overlapping if their summits were within 200 bp of each other. Read mapping statistics and QC metrics for these datasets can be found in Supporting Information, Table S2.

## RESULTS

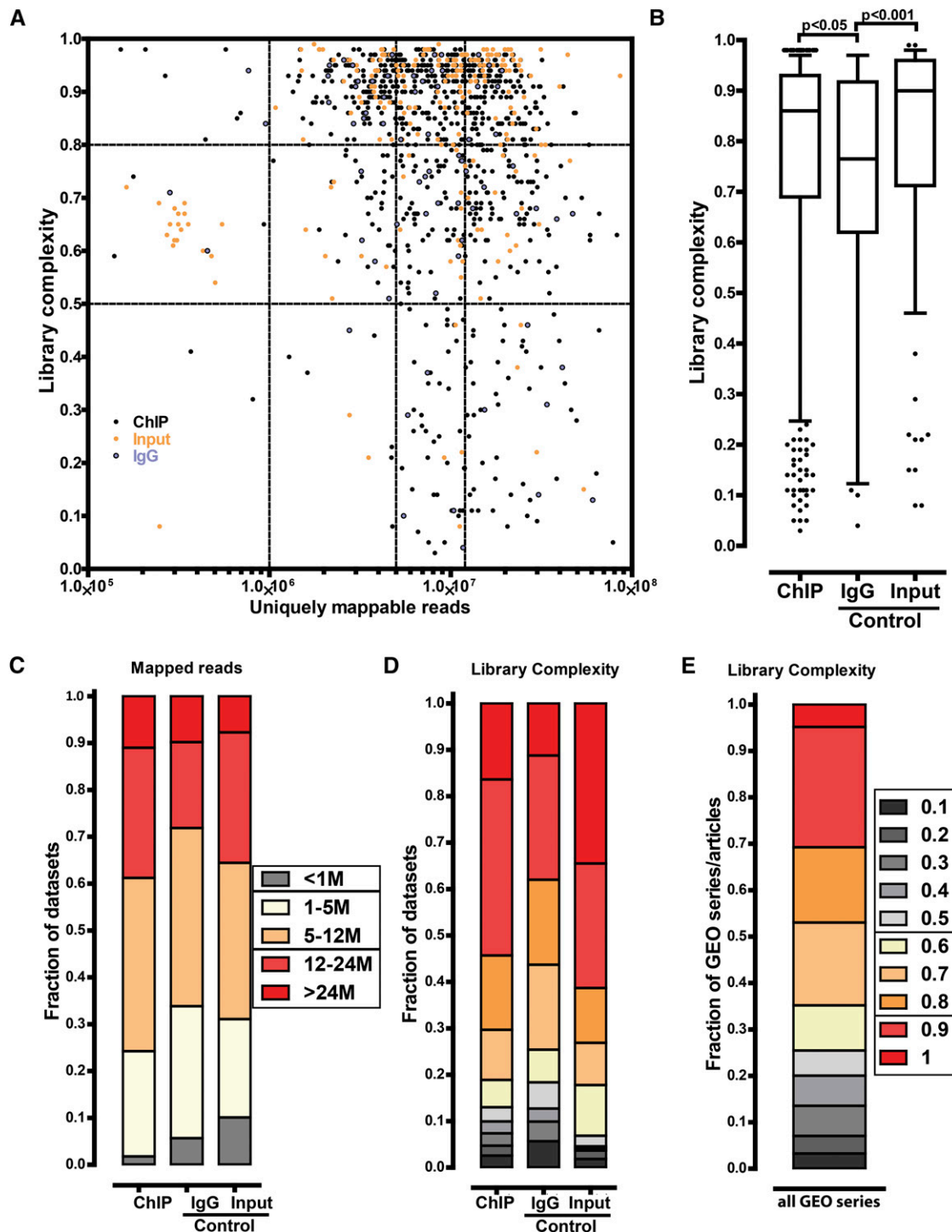
### Dataset collection, data processing, and quality metrics

We downloaded all GEO series containing ChIP-seq datasets for vertebrate TFs or chromatin-modifying and remodeling proteins, along with their corresponding control libraries, submitted before April 1, 2012. We excluded ENCODE datasets because they have previously been subjected to this quality assessment (ENCODE Project Consortium 2012). We provide here a summary of ENCODE TF ChIP-seq data quality from the two main production groups in Figure S9 and Figure S10 (Landt *et al.* 2012).

For several reasons, we also excluded histone modifications and RNA Polymerase II datasets. First, in our experience, ChIP-seq against these targets is very robust to experimental variation and the success rate is reliably high (provided the antibody reagents used are of high quality). Second, an especially large proportion of published data are for histone marks. The effect of including all of these in the survey is to obscure or

skew what is happening in the information-rich sample set that includes diverse TFs and cofactors. Finally, the currently available QC metrics were designed and are best suited for TF data that produce highly localized “point-source” occupancy (as they quantify the extent of read clustering in the genome). This means that the metrics themselves need to be interpreted differently if they are applied to, for example, repressive histone marks such as H3K9me3 and H3K27me3, which form large “broad-source” regions of enrichment (Pepke *et al.* 2009). Arguably, these data will need their own metrics and this will be a challenge for the future.

The final collection of datasets contained 191 GEO series containing a total of 917 ChIP-seq and 292 control libraries. Except for a limited number of cases in which a GEO series was associated with multiple publications, two or three GEO series were associated with the same publication, or a GEO series has not yet been used in a publication, and there is a one-to-one relationship between GEO series and published articles in the literature (Robertson *et al.* 2007; Chen *et al.* 2008; Marson *et al.* 2008; Bilodeau *et al.* 2009; Cheng *et al.* 2009; De Santa *et al.* 2009; Lister *et al.* 2009; Nishiyama *et al.* 2009; Visel *et al.* 2009; Welboren *et al.* 2009; Wilson *et al.* 2009; Yu *et al.* 2009; Yuan *et al.* 2009; Barish *et al.* 2010; Blow *et al.* 2010; Blow *et al.* 2010; Cao *et al.* 2010; Chi *et al.* 2010; Chia *et al.* 2010; Chicas *et al.* 2010; Corbo *et al.* 2010; Cuddapah *et al.* 2009; Durant *et al.* 2010; Fortschegger *et al.* 2010; Gotea *et al.* 2010; Gu *et al.* 2010; Han *et al.* 2010; Heinz *et al.* 2010; Heng *et al.* 2010; Ho *et al.* 2009; Hollenhorst *et al.* 2009; Hu *et al.* 2010; Johannes *et al.* 2010; Jung *et al.* 2010; Kagey *et al.* 2010; Kassouf *et al.* 2010; Kim *et al.* 2010; Kong *et al.* 2010; Kouwenhoven *et al.* 2010; Krebs *et al.* 2010; Kunarso *et al.* 2010; Kwon *et al.* 2009; Law *et al.* 2010; Lee *et al.* 2010; Lefterova *et al.* 2010; Li *et al.* 2010; Lin *et al.* 2010; Liu *et al.* 2010; Ma *et al.* 2010; MacIsaac *et al.* 2010; Mahony *et al.* 2010; Martinez *et al.* 2010; Palii *et al.* 2010; Qi *et al.* 2010; Rada-Iglesias *et al.* 2010; Rahl *et al.* 2010; Ramagopalan *et al.* 2010; Ramos *et al.* 2010; Schlesinger *et al.* 2010; Schnetz *et al.* 2010; Sehat *et al.* 2010; Steger *et al.* 2010; Tallack *et al.* 2010; Tang *et al.* 2010; Vermeulen *et al.* 2010; Verzi *et al.* 2010; Vivar *et al.* 2010; Wei *et al.* 2010; Woodfield *et al.* 2010; Yang *et al.* 2010; Yao *et al.* 2010; Yu *et al.* 2010; An *et al.* 2011; Ang *et al.* 2011; Bergsland *et al.* 2011; Bernt *et al.* 2011; Botcheva *et al.* 2011; Brown *et al.* 2011; Bugge *et al.* 2011; Ceol *et al.* 2011; Ceschin *et al.* 2011; Costessi *et al.* 2011; Ebert *et al.* 2011; Fang *et al.* 2011; Handoko *et al.* 2011; He *et al.* 2011; Heikkinen *et al.* 2011; Holmstrom *et al.* 2011; Horiuchi *et al.* 2011; Hu *et al.* 2011; Joseph *et al.* 2010; Kim *et al.* 2011; Klisch *et al.* 2011; Koeppel *et al.* 2011; Kong *et al.* 2011; Little *et al.* 2011; Liu *et al.* 2011; Lo *et al.* 2011; Marban *et al.* 2011; Mazzoni *et al.* 2011; McManus *et al.* 2011; Mendoza-Parra *et al.* 2011; Meyer *et al.* 2012; Miyazaki *et al.* 2011; Mullen *et al.* 2011; Mullican *et al.* 2011; Nakayamada *et al.* 2011; Nitzsche *et al.* 2011; Norton *et al.* 2011; Novershtern *et al.* 2011; Quenneville *et al.* 2011; Rao *et al.* 2011; Rey *et al.* 2011; Sahu *et al.* 2011; Schmitz *et al.* 2011; Seitz *et al.* 2011; Shen *et al.* 2011; Shukla *et al.* 2011; Siersbæk *et al.* 2011; Smeenk *et al.* 2011; Smith *et al.* 2011; Soccio *et al.* 2011; Stadler *et al.* 2011; Sun *et al.* 2011; Tan *et al.* 2011a; Tan *et al.* 2011b; Teo *et al.* 2011; Tijssen *et al.* 2011; Tiwari *et al.* 2011a; Tiwari *et al.* 2011b; Trompouki *et al.* 2011; van Heeringen *et al.* 2011; Verzi *et al.* 2011; Wang *et al.* 2011a; Wang *et al.* 2011b; Wei *et al.* 2011; Whyte *et al.* 2011; Wu *et al.* 2011a; Wu *et al.* 2011b; Xu *et al.* 2011; Yang *et al.* 2011; Yildirim *et al.* 2011; Yoon *et al.* 2011; Zhang *et al.* 2011; Zhao *et al.* 2011a; Zhao *et al.* 2011b; Avvakumov *et al.* 2012; Barish *et al.* 2012; Boergesen *et al.* 2012; Bugge *et al.* 2012; Canella *et al.* 2012; Cardamone *et al.* 2012; Cheng *et al.* 2012; Chlon *et al.* 2012; Cho *et al.* 2012; Doré *et al.* 2012; Fan *et al.* 2012; Feng *et al.* 2011; Fong *et al.* 2012; Gao *et al.* 2012; Gowher *et al.* 2012; Hunkapiller *et al.* 2012; Hutchins *et al.* 2012; Li



**Figure 1** Sequencing library characteristics. (A) Joint distribution of library complexity and sequencing depth for all datasets examined. Vertical lines are drawn at 1 million, 5 million, and 12 million reads. Horizontal and vertical lines indicate quality classes discussed in the text. The upper right domain (number of uniquely mappable reads  $\geq 12$  million and library complexity  $\geq 0.8$ ) passes current quality thresholds. (B) Distribution of library complexity for ChIP-seq datasets, IgG controls, and inputs. (C) Distribution of sequencing depth for ChIP-seq datasets, IgG controls, and sonicated inputs. (D) Fraction of ChIP-seq, IgG, and input datasets exhibiting high, medium, and low complexity. (E) Fraction of studies containing libraries of high, medium, and low complexity (the distribution of the minimum library complexity observed is shown)

*et al.* 2012; Lu *et al.* 2012; Miller *et al.* 2011; Ntziachristos *et al.* 2012; Pehkonen *et al.* 2012; Ptasinska *et al.* 2012; Remeseiro *et al.* 2012; Sadasivam *et al.* 2012; Sakabe *et al.* 2012; Schödel *et al.* 2012; Trowbridge *et al.* 2012; Vilagos *et al.* 2012; Wu *et al.* 2012; Xiao *et al.* 2012; Yu

*et al.* 2012; unpublished at the time of completion of this manuscript are the following GEO accession numbers: GSE33346, GSE33850, GSE36561, GSE30919, GSE33128, GSE35109, GSE25426, GSE31951, GSE26711, GSE23581, GSE26136, GSE26680, GSE15844, GSE21916,



GSE22303, and GSE29180; direct links to all GEO series can be found in Table S1).

We discuss IgG and input controls separately because, to the best of our knowledge, any potential general differences between the two types of controls have not been investigated systematically in the context of ChIP-seq (Peng *et al.* 2007 addressed these questions for ChIP-Chip data; however, the nature of the background is substantially different for microarrays).

We mapped all reads with uniform settings (see *Materials and Methods* for details) and examined library and ChIP QC metrics for each dataset. These criteria have already been discussed by Landt *et al.* (2012), and a detailed treatment of cross-correlation is presented elsewhere (Kundaje *et al.*, unpublished data). Here, we provide a brief overview of each.

**Sequencing depth:** If a ChIP-seq experiment achieves successful immune enrichment and the resulting library adequately represents the sample, then greater sequencing depth will produce a more complete map of TF occupancy (Landt *et al.* 2012). At a greater depth, the measurement will identify a larger number of reproducible sites containing the corresponding DNA-binding sequence motif. Under-sequencing of an otherwise successful library will lead to false-negative results. It has been difficult to establish a universal minimal sequencing depth because of differences between factors. Any threshold is going to be somewhat arbitrary but, in general, the major cost/benefit trade-off is between sequencing individual samples more deeply and generating more replicates; for most contemporary purposes, an independent duplicate measurement of 12 million reads arguably adds greater overall value than a single determination with 24 million reads, even though the higher number of reads will increase sensitivity. The number of mapped reads less than 1–2 million for a typical TF will

usually be inadequate for capturing the complexity of an interactome for a mammalian-size genome. Many datasets now in the public domain were generated when sequencing throughput was lower than it is now and costs were higher (between 2007 and 2013, sequencing throughput has increased by approximately two orders of magnitude). As a consequence, many early ChIP-seq libraries were sequenced to a depth of only a few million reads. We therefore divided datasets into sequencing bins by using thresholds of 1 million, 5 million, 12 million, and 24 million uniquely mapped reads (taking into account sequencing depths recommended in the past by the ENCODE consortium for TFs). Libraries having less than 1 million reads are considered severely undersequenced, and those with more than 12 million are considered reasonably deeply sequenced.

**Library complexity:** A second characteristic that influences the quality of a ChIP-seq measurement is the sequence fragment diversity of the sequencing library. This is often referred to as library complexity, and low complexity is undesirable, although we note that much better IP enrichment than what is now obtained could, in the future, lead to very high-quality datasets with low library complexity. Currently, low-complexity libraries mainly result from experimental deficiencies: either too few starting molecules at the end of the immunoprecipitation step or inefficient steps in subsequent library building. As a result, the same starting molecules are sequenced repeatedly. Very-low-complexity libraries will not contain enough information to effectively sample the true positive occupancy sites and they distort the signal position and intensity. This can confuse peak callers (especially if the algorithm does not collapse presumptive PCR duplicates), leading to peak calling artifacts (Landt *et al.* 2012). We calculate the following metric as an indicator of library complexity (Landt *et al.* 2012):

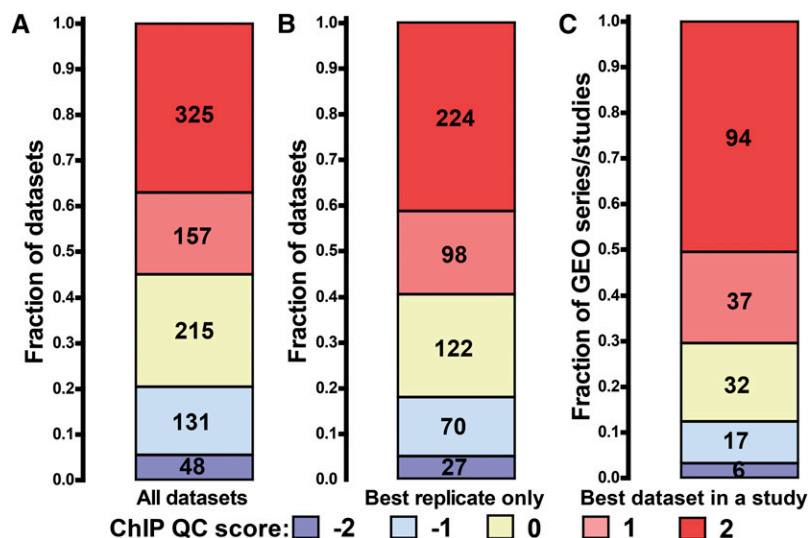
$$\text{Library complexity} = \frac{\text{Number positions in the genome with uniquely mappable reads in dataset}}{\text{Number uniquely mappable reads in dataset}} \quad (1)$$

Estimated in this simple way, library complexity is expected to decrease eventually with increased sequencing depth because even highly complex libraries become exhausted by very deep sequencing. Reduced apparent complexity would also be observed with extremely successful ChIP-seq experiments for TFs that bind to the genome in a highly discriminative fashion to a limited number of locations. In such libraries, the majority of reads would originate from the limited genomic subspace around binding sites, resulting in low library complexity. With current methods, this is a largely theoretical consideration; in practice, in most ChIP-seq libraries only a minority of reads originates from factor-bound sites, with the rest (the majority) representing genomic background. Because the majority of libraries we examined were in the sequencing depth range over which these values represent library complexity reasonably well (Figure 1A and Figure S2), we separated datasets into the following complexity groups: high complexity (apparent library complexity  $\geq 0.8$ ); medium to low complexity (apparent library complexity between 0.5 and 0.8); and very low complexity (apparent library complexity  $\leq 0.5$ ). We also note that in substantially smaller genomes, the apparent library complexity is expected to be lower because the number of positions from which sequencing library fragments can originate is smaller.

**Cross-correlation analysis of read clustering and ChIP enrichment:** Because the majority of sequencing reads in a ChIP-seq library

represent nonspecific genomic backgrounds, these reads are expected to be distributed randomly over the genome, to a first approximation. In contrast, reads originating from specific occupancy events cluster around the sites of protein–DNA interactions, where they are distributed in a characteristic asymmetric pattern on the plus and minus strands (Kharchenko *et al.* 2008). Cross-correlation analysis is an effective way of measuring the extent of this clustering. It also captures additional global features of the data, such as the average fragment length and fragment length distribution (Kharchenko *et al.* 2008; Landt *et al.* 2012). Specifically, the read coverage profiles on the two strands are shifted relative to the other over a range of shift values and the correlation between the profiles is calculated at each shift (Kharchenko *et al.* 2008). The resulting plot has one (“phantom”) peak corresponding to the read length and another peak corresponding to the average fragment length; the height of the fragment-length peak is highly informative of the extent of read clustering in the library and, in turn, of the success of a ChIP-seq experiment. This feature is best captured by the NSC and RSC metrics discussed by Landt *et al.* (2012).

We applied SPP (Kharchenko *et al.* 2008) to perform cross-correlation analysis for all libraries in our survey. We then used the RSC cross-correlation metric to assign integer QC tag values in the  $\{-2, 2\}$  range to datasets, with QC values of 2 corresponding to very highly clustered (and most likely, also successful) datasets and QC values of  $-2$  to datasets exhibiting no to minimal read clustering; negative values are



**Figure 2** ChIP QC assessment summary. The numbers in each box indicate the total number of datasets/studies belonging to it. SPP QC scores of +1 and +2 indicate a high degree of read clustering in a dataset. (A) Distribution of SPP QC scores for all ChIP-seq datasets examined. (B) Distribution of SPP QC scores for the best replicates for a factor/condition combination in each study. (C) Distribution of the maximum SPP QC scores for all ChIP-seq datasets in a study.

expected for input datasets. The RSC metric captures well the extent of read enrichment in vertebrate genomes similar in size and structure to humans, which this study focuses on. We provide representative examples of cross-correlation plots for each of the five QC categories in Figure S1A, and we use these tags as convenient general proxies for ChIP quality throughout the following analysis. We note that the discretization thresholds are not intended to be absolute determinants of quality, but they do enable one to rapidly scan very large numbers of datasets. In practice, examining the cross-correlation plots and the continuously distributed NSC and RSC values and using those together with information about sequencing depth and library complexity are always more informative and can provide valuable nuances for understanding specific datasets. Direct examination of plots allows one to detect datasets with odd cross-correlation profiles (we show a few representative examples in Figure S11). It is possible in theory for low-complexity libraries to produce artificially high cross-correlation scores if stacks of reads on opposite strands are located close to each other in regions of enrichment; however, the Pearson correlation between library complexity scores and RSC values in the collection of ChIP datasets surveyed here was 0.0084, indicating that such cases do not feature significantly in this analysis.

An additional major component of the ChIP-seq QC pipeline developed by the ENCODE consortium is reproducibility analysis of replicates, based on the irreproducible discovery rate (IDR) statistic (Li *et al.* 2011). However, because many of the studies we surveyed did not have replicates, we only evaluated datasets on the level of individual experiments. Single dataset evaluation is almost always a valuable precursor to evaluation of replicates because, typically, a second replicate is generated after a successful first one. The full list of datasets, mapping, and QC statistics is provided in Table S1.

### Sequencing depth and library complexity

Figure 1A shows the distribution of sequencing depth and library complexity for ChIP-seq and control datasets. The upper right domain, bounded by 12 million reads per sample and a complexity value of 0.8, is an arbitrary but useful definition of high quality according to these measures. A majority of datasets had reasonably good complexity and severely undersequenced libraries were rare (Figure 1C). A minority (38.8%) of datasets had more than 12 million mapped reads; however, as discussed, this is not unexpected, because a large fraction of the datasets we surveyed were generated in times of sig-

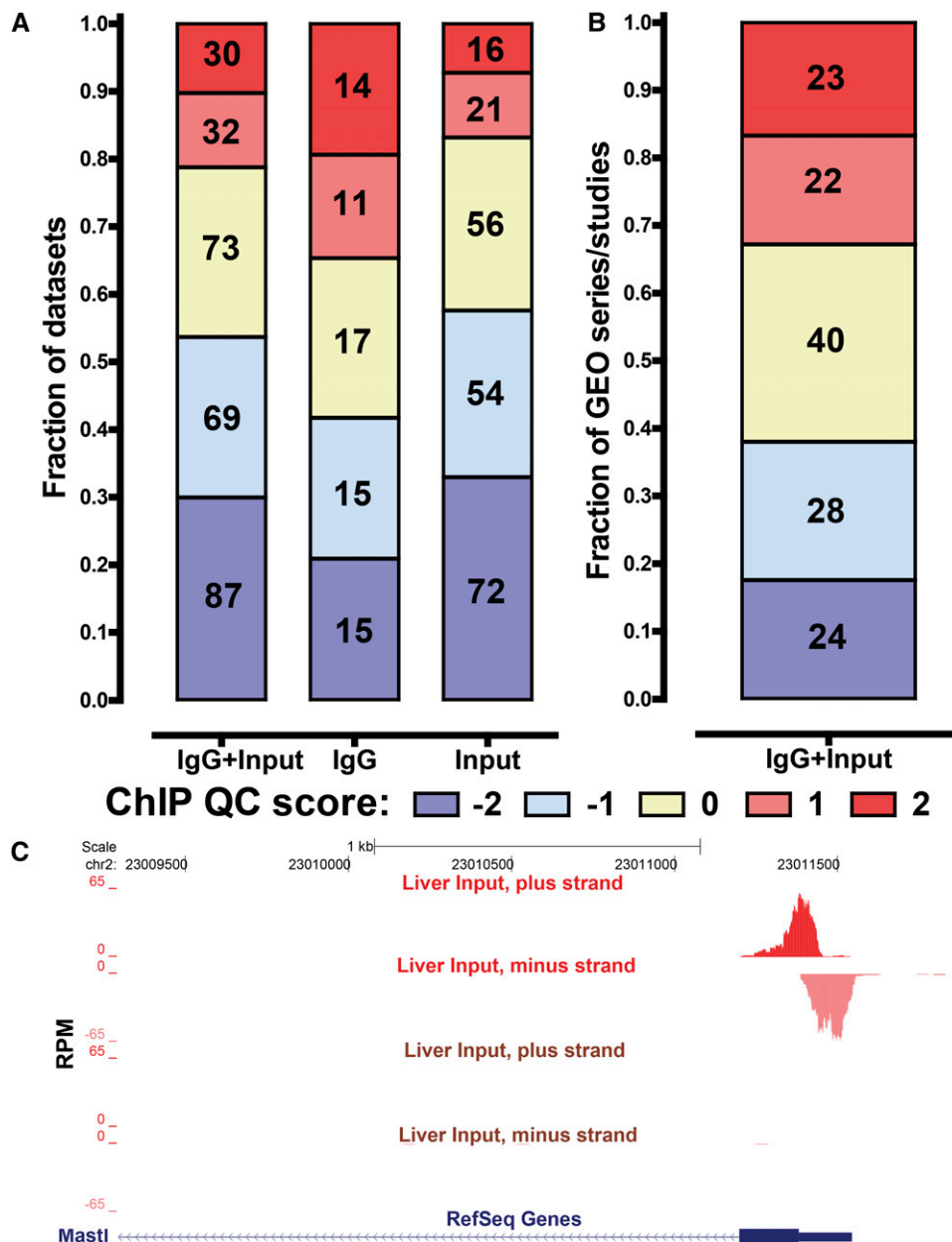
nificantly higher sequencing cost and lower throughput. Strikingly, the median complexity of IgG control datasets was less than 0.8 and considerably lower than that of either ChIP-seq or sonicated input libraries (Figure 1B). This is not a result of IgG datasets having been sequenced much more deeply than the other two groups; in fact, the median sequencing depth of IgG controls is lower (Figure S2). The concern that some individual IgG inputs might provide insufficient DNA mass to build highly complex libraries has been raised before (Landt *et al.* 2012), and our observations are consistent with this, although it is not a characteristic of all IgG controls.

Slightly more than half (54.3%) of ChIP-seq datasets had library complexity more than 0.8, whereas very-low-complexity (< 0.5) libraries comprised 12.9% of datasets; the fraction of very-low-complexity libraries was higher and lower for IgG and input datasets, respectively (Figure 1D). Because most GEO series contained multiple libraries, we also asked, how common is the presence of low-complexity libraries in individual studies? Figure 1E shows the distribution of the minimum library complexity in each such series (for all types of datasets). One-quarter (25.4%) of all studies contained very-low-complexity libraries.

### Cross-correlation quality assessment of ChIP-seq datasets

Next, we examined the distribution of SPP QC scores for ChIP-seq datasets. Before doing this, we excluded a minority of datasets for which there was a good reason to think high ChIP enrichment should not be expected. For example, experiments executed in knockouts, knockdowns, or settings in which the factor is not expressed are not expected to produce a high-scoring measurement. And in a few cases, the factor in question might be known to bind to only a small number of sites in the genome; this has been proposed, for example, for some ZNF TFs and Pol3 and its associated factors (Landt *et al.* 2012). Our detailed criteria for inclusion are described in *Materials and Methods*.

Figure 2A shows the QC score distribution for all ChIP-seq datasets we retained. Strikingly, only 55% (482 out of 876) of datasets had QC scores of 1 or 2, *i.e.*, they were likely to be highly successful. An additional 24.5% (215 out of 876) had a score of 0, indicating that they were of intermediate quality, and 20.4% (179 out of 876) had low-quality scores of -1 and -2. Sometimes multiple replicates for a factor were submitted but only one scored poorly, so we also compiled a second set of ChIP-seq experiments that only included the best available replicate for each factor and condition (Figure 2B). This



**Figure 3** Assessment of read clustering in control datasets. The numbers in each box indicate the total number of datasets/studies belonging to it. SPP QC scores of 1 and 2 indicate a high degree of read clustering in a dataset. (A) Distribution of SPP QC scores for all control datasets (IgG + input), IgG/mock IP controls (IgG), and sonicated inputs (inputs). (B) Fraction of studies containing highly clustered inputs. The distribution of the maximum SPP QC score for all inputs in a dataset is shown. (C) Examples of a highly clustered input [mouse liver, upper two tracks, (MacIsaac *et al.* 2010), QC score of 2] and an input that does not show high extent of read clustering [mouse liver, lower two tracks (Soccio *et al.* 2011), QC score of -1]. The promoter of the *MASTL* gene is shown. All tracks are shown to the same scale and reads mapping to the plus and minus strands are displayed separately for better visualization of the cross-correlation between the two.

set included 322 datasets (59%) with QC scores of 2 or 1. The fraction of intermediate-quality or low-scoring datasets in this group decreased as expected. However, the decrease was modest with 18% (97 out of 541) of the best available replicates scoring -1 or -2, and 22.5% (122 out of 541) scoring 0.

We then examined the distribution of the maximum QC score for each study, regardless of the target identity (Figure 3C). The fraction of low scores decreased further, though only 70.4% of studies (131 out of 186) had a score of 1 or 2 for their best experiment. Finally, we compiled a list of the top-scoring datasets from all studies that assayed only a single TF; 19.7% (19 out of 96) of these studies had scores of -1 or -2, 25% (24 of 96) had a score of 0, and 55.2% (53 of 96) were marked as likely to be successful, with scores of 1 and 2 (Figure S3C).

### Read clustering in control datasets

Control datasets serve the important purpose of helping to distinguish read enrichment attributable to the immunoprecipitation step from

artificial read clustering attributable to other experimental factors, both known and unknown. It is, for example, well-appreciated that differential chromatin shearing efficiency can lead to the overrepresentation of areas of open chromatin (usually immediately surrounding transcribed promoters) in sequencing libraries. This has been termed the “Sono-seq” effect when attributed to sonication (Auerbach *et al.* 2009). In addition, unknown copy number variants relative to the reference genome or sequence composition biases can give false-positive occupancy calls. In particular, specifics of the amplification step in sequencing platforms can introduce bias due to GC content (Ho *et al.* 2011).

In general, control datasets are not expected to exhibit a pattern of significant read clustering similar in strength to that of successful ChIP-seq datasets. In our own practice, under standard cross-linking protocols, most do not. However, we noticed that a minority of control datasets produce positive ChIP QC metric scores along with prominent cross-correlation peaks. Figure S1B shows examples of cross-correlation plots for individual control datasets with all possible QC scores, from



−2 to 2, and Figure 3C shows a browser snapshot of a region with strong read enrichment in a highly clustered (QC score of 2) input library. No such enrichment was observed in a different control library from a similar biological source having a QC score of −1.

We asked how general this phenomenon is by examining the distribution of QC scores of both IgG and input control datasets (Figure 3A). Surprisingly, only 53.6% (156 out of 291) of control datasets had QC scores of −2 or −1 and 25% (73 of 291) had a score of 0, whereas 21.3% (62 of 291) exhibited a very high degree of read clustering and received scores of 1 or 2. The highly clustered inputs were notably more common among IgG controls than among input chromatin controls (Figure 3A). Moreover, high read clustering was more often found in low-complexity libraries (which are themselves more common among IgG controls) (Figure S4, A and B).

We also examined how widespread control sample clustering is on the level of individual GEO series/studies to see if the phenomenon is restricted to a few larger studies. Figure 3B shows the distribution of the maximal control sample QC score for all studies. Of the studies for which control datasets were available, 32.8% (45 of 123) contained at least one highly clustered control with a score of 1 or 2, and 29.2% (40 of 123) contained a control with a score of 0. Thus, control datasets surprisingly often exhibit a high extent of read clustering similar to that of ChIP-seq datasets. This is even more striking considering that formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) data (an assay that is based on the preferential enrichment of open chromatin in sonicated DNA and aims to achieve high read clustering) from ENCODE usually have QC scores between −2 and 0. Moreover, the Sono-seq datasets published by Auerbach *et al.* (2009) all have scores of −2.

We note that unless this effect is very strong and is associated with notable genomic features such as promoters of genes, it can be difficult to detect by the usual methods of visual inspection of signal tracks on a genome browser. It is, however, readily apparent in cross-correlation analysis and our results raise awareness of its existence. As mentioned, one candidate explanation for this phenomenon is the previously described “Sono-seq” effect. Using standard experimental protocols, this effect has been rare in our experience; however, under more aggressive cross-linking conditions, we have observed increased read clustering in control samples (Figure S5). Notably, the original “Sono-seq” description focused on promoter regions, but we have also observed it over distal regulatory elements, where its strength was even higher than at promoters (Figure S5). Thus, variation in the extent of fixation, as well as sonication, might be a substantial contributor to variation in read clustering across the broader data collection. Another potential contributing factor is sequencing depth. Although the average sequencing depth for highly clustered IgG and input controls is higher than that of controls with negative QC scores (Figure S4, C and D) this by no means explains all the clustering observed in controls. There are many examples of more deeply sequenced input and IgG libraries with no significant cross-correlation peaks and very few of them were sequenced especially deeply (only eight control libraries had  $>4 \times 10^7$  reads not desirable). Finally, “Sono-seq” need not be the only explanation. Whereas a number of control datasets with QC scores of 2 exhibited higher read coverage around promoters, others did not (Figure S6), suggesting at least one additional source of unexplained read enrichment in control samples. Because rich annotation of functional genomic elements outside promoter regions was not available for many cell types in our survey, this phenomenon is a subject for future analyses.

## DISCUSSION

We performed a systematic survey of ChIP quality for publicly available vertebrate ChIP-seq datasets and found that more than half

score as high quality by our measures. This group comprises a set that we believe can be used with confidence for integrative analyses. This conclusion carries the important caveat that we could not assess the specificity of the immune reagents used to perform the experiments, which powerfully affects the biological meaning of the data.

A substantial minority of published datasets (between 20% and 45% of those examined) were of low or intermediate quality by our metrics. This was true not only for individual libraries but also for the best replicates from each study. In addition, we observed a substantial number of low-complexity datasets and an unexpected group of highly clustered control datasets. These observations underscore the widespread variation in published ChIP-seq data. They also raised questions about which kinds of conclusions in primary publications are more or less sensitive to these aspects of data quality. In particular, global quality analysis is useful for guiding subsequent re-use of published data that require higher quality than was needed or achieved in the source study.

Data quality varied widely across “impact” levels. We separated datasets into groups according to the 2011 Thomson Reuters Impact Factor for the journal in which the corresponding article was published and examined the distribution of QC scores in each group (Figure S8). The group with highest impact factor ( $\geq 25$ ) contained the largest fraction of datasets with a low QC score of −2 or −1. We also examined the distribution of QC scores with respect to the year of publication and found that the fraction of datasets with low scores has stabilized in the past 3 yr at approximately 20% (Figure S7).

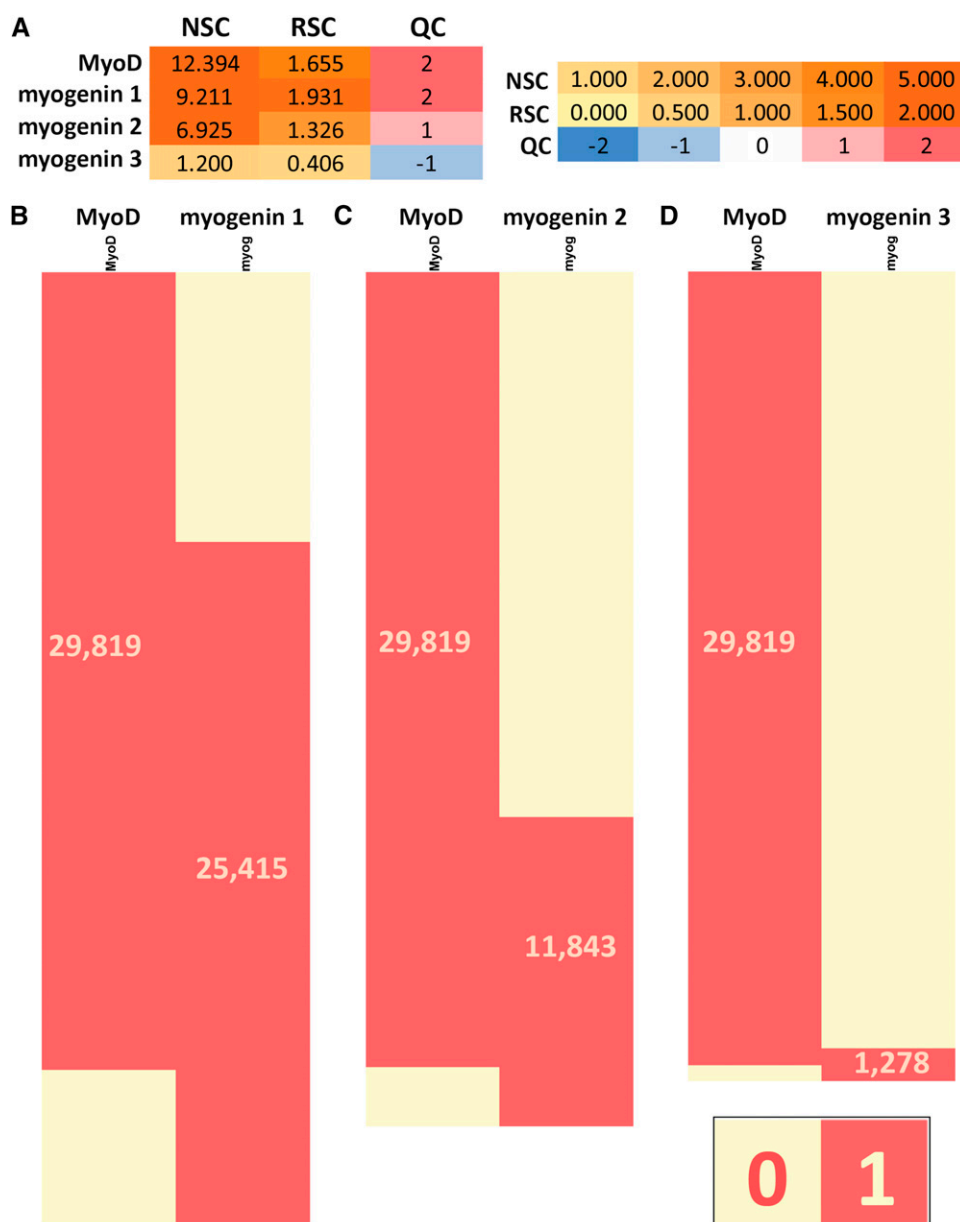
We emphasize that datasets scoring as low quality by the metrics used here can, nevertheless, produce important biological discoveries. For this reason, it would be an error to set a rigid “standard” that every published dataset must meet. Instead, routine QC analysis can make it easy to see when there is reason for concern about a given dataset. It can also provide a first tier of guidance about what uses are likely to be appropriate for a given dataset. As discussed previously, the appropriate level of QC stringency depends on the specific goals of the experiment and methods of analysis (Landt *et al.* 2012). In particular, some analyses that are sensitive to false-negative results are particularly vulnerable to inclusion of low-scoring datasets. For example, trying to derive combinatorial TF occupancy rules is seriously compromised and even misleading if a subset of the datasets included is suboptimal.

We illustrate this with a simple example from our own experience (Figure 4). The MyoD and myogenin TFs are well-known regulators of muscle differentiation (Yun and Wold 1996) and C2C12 cells (Yaffe and Saxel 1977) have been widely used to study the process because they can be propagated in an undifferentiated myoblast state and easily induced to differentiate into myocytes and myotubes. We have performed several ChIP-seq experiments with these factors in differentiated and undifferentiated C2C12 cells (G. DeSalvo *et al.*, unpublished data; A. Kirilusha *et al.*, unpublished data; K. Fisher-Aylor *et al.*, unpublished data), some of which have been highly successful, whereas others were of poor or intermediate quality. Here, we examined the effect of weaker ChIP-seq datasets on combinatorial occupancy analysis using a MyoD ChIP-seq dataset with very high QC metrics and three myogenin datasets with very high, moderately good, and very low metrics (Figure 4A). Using the best myogenin dataset, we found a high degree of overlap between the binding sites of the two factors (Figure 4B). When the medium-quality myogenin dataset was used instead, a sizable group of MyoD-only sites emerged (Figure 4C) and the erroneous conclusion that a substantial number of MyoD sites lack myogenin binding could be reached if this was the only dataset available for analysis. Finally, the poor-quality myogenin dataset contains very few called peaks and, as a result, almost all MyoD sites show no myogenin binding when it is used for analysis (Figure 4D).

Recently, IDR analysis of replicate datasets (Li *et al.* 2011; ENCODE Project Consortium 2012; Landt *et al.* 2012) emerged as a robust method for deriving lists of reproducible occupancy sites from ChIP-seq datasets. IDR is based on differences in the consistency of ranking (usually by signal strength as measured by read enrichment or by statistical significance) for all identified peaks in a pair of ChIP-seq replicates. A virtue of this approach is that it allows a statistically robust set of binding sites to be derived largely independent of thresholds and settings specific to a particular peak-calling algorithm. Ideally, IDR would be used in conjunction with the quality metrics used here (ENCODE Project Consortium 2012; Landt *et al.* 2012). However, replicate measurements do not exist for many of the datasets in our survey of the historic. We expect that IDR will become common practice as sequencing costs decline. Even when that happens, measurements of the quality of individual datasets will remain important because they capture specific information in addition to reproducibility and because IDR analysis is sensitive to the presence of poor-

quality replicates. An asymmetric pair consisting of one high-quality and one poorer-quality dataset is dominated in IDR by the weaker replicate, resulting in a shorter list of sites and a high false-negative rate. Care should be exercised in such cases. Although the best approach is to obtain a second high-quality replicate, but if this is not possible, special strategies for treating asymmetric replicates have been devised (Landt *et al.* 2012).

The most perplexing observation was that a subset of control datasets have extensive read clustering in the same range as successful ChIP-seq experiments. In our own practice, we have rarely encountered such libraries and, to the best of our knowledge, there has been no extensive treatment of this issue or its influence on data analysis in the literature. The phenomenon occurred more frequently in IgG controls than in input chromatin controls, although it is by no means limited to the former. In theory, an IgG control should be a superior representation of the true background noise in a ChIP-seq sample because it incorporates biases introduced by the entire



**Figure 4** Effect of suboptimal datasets on combinatorial occupancy analysis. The muscle-regulatory factors MyoD and myogenin were assayed in C2C12 myocytes at 60 hr after differentiation. Shown are a single, highly successful MyoD ChIP-seq dataset and three myogenin ChIP-seq datasets, one of which is similarly highly successful ("myogenin 1"), a second weaker one ("myogenin 2"), and a third one that is an experimental failure ("myogenin 3"). (A) Quality control metrics. (B, C, D) The extent of overlap of MyoD and myogenin-binding sites as determined using each of the three myogenin datasets (see *Materials and Methods* for data processing details). MyoD and myogenin are mostly found to bind to the same sites when inter-actome determinations of comparable strength are used. (B) A sizable group of apparently MyoD-only sites emerges when the medium-strength myogenin dataset is used because of a large number of false-negative myogenin calls. (C) Finally, the unsuccessful myogenin ChIP reveals that most MyoD are not shared by myogenin. (D) Numbers listed in the red blocks corresponding to each set of peak calls indicate size.

immunoprecipitation process, in addition to any enrichments or biases created by chromatin shearing. Using this logic, a simple interpretation is that high read clustering in these controls correctly identifies artifacts in the IP process. When high background sample clustering is observed in control sample, we suggest that it merits immediate investigation of its replicability and its impact on peak-calling for the corresponding ChIP-seq samples. The fact that we also observed a large number of IgG controls (Figure 3A) that showed no such clustering, argues that this is not a general feature.

A crucial issue is the extent to which clustering in controls is also present as experimental noise in ChIP-seq libraries from the same material. In other words, how well-matched are the control samples with the corresponding experimental samples, and how robust are the controls? For example, a very strong Sono-seq effect in a control sample is expected to give ChIP-seq libraries with high read clustering that is a combination of true ChIP (antibody-specific) signal plus Sono-seq-derived noise that covers promoters and enhancers in a non-specific manner. Whereas most contemporary peak callers normalize for enrichment in controls, very strong background noise will diminish the signal-to-noise ratio and adversely affect sensitivity. How severely this affects the results will depend on the overlap between true factor occupancy sites and regions of artifactual read enrichment (for some factors this overlap may be negligible because they do not bind to Sono-seq regions); on the magnitude of the Sono-seq effect; and on the strength of the ChIP itself (sufficiently strong determinations are not greatly affected). Conversely, if a ChIP-seq library has a strong Sono-seq component and peak calling is performed against an imperfectly matched “control” sample in which the Sono-seq effect is of significantly lower magnitude, false-positive peak calls will increase. Unfortunately, in practice such cases are difficult to detect. They are not flagged directly by current quality metrics and are best detected by analyses specific to each study and factor, including specific motif enrichment, especially when little is known about the expected true-positive rates. Similar reasoning applies if the noise source is something other than Sono-seq.

Uniform retrospective quality assessment is resource-intensive and will not be practically feasible because the number of ChIP-seq datasets is growing exponentially. Retrospective analysis also comes too late to influence the experiments themselves or to contribute to the review process. A reasonable path forward would be to incorporate routine data quality assessment into experimental analysis, review for publication, and submission to public repositories, as a matter of community practice. However, our results also strongly caution against the blind and arbitrary application of our metrics (or others) in the absence of experimental and biological context. The character of the metrics used here reflects contemporary technology and the quality scale has been calibrated based on factors and co-factors most studied to date. We have seen that it is possible for good datasets to receive low QC scores in certain special situations (*e.g.*, very few sites of occupancy in the genome). It is also possible for some poor or mediocre datasets to receive high QC scores. For example, this can happen as a side-product of strongly clustered backgrounds of the kind discussed above. Some examples of datasets in which this might be the case are shown in Figure S11. For factors that ChIP extremely well, even datasets that are substantially suboptimal score highly. For example, CTCF ChIP-seq datasets routinely identify 35,000–40,000 reproducible binding sites and have QC scores of 2; a dataset that identifies only 15,000 sites is suboptimal given that knowledge; yet it will still receive a positive QC score. For these reasons, the current quality metrics are best used in the context of what is known about the factor, the biological system, and the questions being asked.

Despite important nuances of interpretation, we suggest that using ChIP quality metrics and making the results readily accessible will facilitate better-informed data use by the wider community. An important adjunct to routine QC annotation would be the ability, in major public data repositories, to flag and explain the exceptional cases for which QC scores should not be taken at face value. Finally, quality metrics themselves will continue to improve as the field’s understanding of data structure, experimental artifacts, and the underlying biology all become more sophisticated. Provisions will be needed for incorporating such advances into routine dataset annotation while still achieving comparability through time.

## ACKNOWLEDGMENTS

We thank members of the ENCODE consortium and members of the Wold laboratory for helpful discussions, and Henry Amrhein, Diane Trout, and Sean Upchurch for computational assistance. G.K.M. and B.J.W. are supported by the Beckman Foundation, the Donald Bren Endowment, and National Institutes of Health grants U54 HG004576 and U54 HG006998.

## LITERATURE CITED

- An, C. I., Y. Dong, and N. Hagiwara, 2011 Genome-wide mapping of Sox6 binding sites in skeletal muscle reveals both direct and indirect regulation of muscle terminal differentiation by Sox6. *BMC Dev. Biol.* 11: 59.
- Ang, Y. S., S. Y. Tsai, D. F. Lee, J. Monk, J. Su *et al.*, 2011 Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* 145: 183–197.
- Auerbach, R. K., G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi *et al.*, 2009 Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. USA* 106: 14926–14931.
- Avvakumov, N., M. E. Lalonde, N. Saksouk, E. Paquet, K. C. Glass *et al.*, 2012 Conserved molecular interactions within the HBO1 acetyltransferase complexes regulate cell proliferation. *Mol. Cell. Biol.* 32: 689–703.
- Barish, G. D., R. T. Yu, M. Karunasiri, C. B. Ocampo, J. Dixon *et al.*, 2010 Bcl-6 and NF- $\kappa$ B cistromes mediate opposing regulation of the innate immune response. *Genes Dev.* 24: 2760–2765.
- Barish, G. D., R. T. Yu, M. S. Karunasiri, D. Becerra, J. Kim *et al.*, 2012 The Bcl6-SMRT/NCoR cistrome represses inflammation to attenuate atherosclerosis. *Cell Metab.* 15: 554–562.
- Barski, A., S. Cuddapah, K. Cui, T. Roh, D. E. Schones *et al.*, 2007 High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823837.
- Bergslund, M., D. Ramsköld, C. Zaouter, S. Klum, R. Sandberg *et al.*, 2011 Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.* 25: 2453–2464.
- Bernt, K. M., N. Zhu, A. U. Sinha, S. Vempati, J. Faber *et al.*, 2011 MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. *Cancer Cell* 20: 66–78.
- Bilodeau, S., M. H. Kagey, G. M. Frampton, P. B. Rahl, and R. A. Young, 2009 SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev.* 23: 2484–2489.
- Blow, M. J., D. J. McCulley, Z. Li, T. Zhang, J. A. Akiyama *et al.*, 2010 ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* 42: 806–810.
- Boergesen, M., T. Å. Pedersen, B. Gross, S. J. van Heeringen, D. Hagenbeek *et al.*, 2012 Genome-wide profiling of liver X receptor, retinoid X receptor, and peroxisome proliferator-activated receptor  $\alpha$  in mouse liver reveals extensive sharing of binding sites. *Mol. Cell. Biol.* 32: 852–867.
- Botcheva, K., S. R. McCorkle, W. R. McCombie, J. J. Dunn, C. W. Anderson *et al.*, 2011 Distinct p53 genomic binding patterns in normal and cancer-derived human cells. *Cell Cycle* 10: 4237–4249.
- Brown, S., A. Teo, S. Pauklin, N. Hannan, C. H. Cho *et al.*, 2011 Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. *Stem Cells* 29: 1176–1185.

- Bugge, A., D. Feng, L. J. Everett, E. R. Briggs, S. E. Mullican *et al.*, 2011 Rev-erb $\alpha$  and Rev-erb $\beta$  coordinately protect the circadian clock and normal metabolic function. *Genes Dev.* 26: 657–667.
- Canella, D., D. Bernasconi, F. Gilardi, G. LeMartelot, E. Migliavacca *et al.*, 2012 A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. *Genome Res.* 22: 666–680.
- Cao, L., Y. Yu, S. Bilke, R. L. Walker, L. H. Mayeenuddin *et al.*, 2010 Genome-wide identification of PAX3-FKHR binding sites in rhabdomyosarcoma reveals candidate target genes important for development and cancer. *Cancer Res.* 70: 6497–6508.
- Cardamone, M. D., A. Krones, B. Tanasa, H. Taylor, L. Ricci *et al.*, 2012 A protective strategy against hyperinflammatory responses requiring the nontranscriptional actions of GPS2. *Mol. Cell* 46: 91–104.
- Ceol, C. J., Y. Houvras, J. Jane-Valbuena, S. Bilodeau, D. A. Orlando *et al.*, 2011 The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* 471: 513–517.
- Ceschin, D. G., M. Walia, S. S. Wenk, C. Duboé, C. Gaudon *et al.*, 2011 Methylation specifies distinct estrogen-induced binding site repertoires of CBP to chromatin. *Genes Dev.* 25: 1132–1146.
- Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss *et al.*, 2008 Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133: 1106–1117.
- Cheng, Y., W. Wu, S. A. Kumar, D. Yu, W. Deng *et al.*, 2009 Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.* 19: 2172–2184.
- Cheng, B., T. Li, P. B. Rahl, T. E. Adamson, N. B. Loudas *et al.*, 2012 Functional association of Gdown1 with RNA polymerase II poised on human genes. *Mol. Cell* 45: 38–50.
- Chi, P., Y. Chen, L. Zhang, X. Guo, J. Wongvipat *et al.*, 2010 ETV1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours. *Nature* 467: 849–853.
- Chia, N. Y., Y. S. Chan, B. Feng, X. Lu, Y. L. Orlov *et al.*, 2010 A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* 468: 316–320.
- Chicas, A., X. Wang, C. Zhang, M. McCurrach, Z. Zhao *et al.*, 2010 Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. *Cancer Cell* 17: 376–387.
- Chlon, T. M., L. C. Doré, and J. D. Crispino, 2012 Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. *Mol. Cell* 47: 608–621.
- Cho, H., X. Zhao, M. Hatori, R. T. Yu, G. D. Barish *et al.*, 2012 Regulation of circadian behaviour and metabolism by REV-ERB- $\alpha$  and REV-ERB- $\beta$ . *Nature* 485: 123–127.
- Corbo, J. C., K. A. Lawrence, M. Karlstetter, C. A. Myers, M. Abdelaziz *et al.*, 2010 CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res.* 20: 1512–1525.
- Costessi, A., N. Mahrour, E. Tijchon, R. Stunnenberg, M. A. Stoel *et al.*, 2011 The tumour antigen PRAME is a subunit of a Cul2 ubiquitin ligase and associates with active NFY promoters. *EMBO J.* 30: 3786–3798.
- Cuddapah, S., R. Jothi, D. E. Schones, T. Y. Roh, K. Cui *et al.*, 2009 Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* 19: 24–32.
- De Santa, F., V. Narang, Z. H. Yap, B. K. Tusi, T. Burgold *et al.*, 2009 Jmjd3 contributes to the control of gene expression in LPS-activated macrophages. *EMBO J.* 28: 3341–3352.
- Doré, L. C., T. M. Chlon, C. D. Brown, K. P. White, and J. D. Crispino, 2012 Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood* 119: 3724–3733.
- Durant, L., W. T. Watford, H. L. Ramos, A. Laurence, G. Vahedi *et al.*, 2010 Diverse targets of the transcription factor STAT3 contribute to T cell pathogenicity and homeostasis. *Immunity* 32: 605–615.
- Ebert, A., S. McManus, H. Tagoh, J. Medvedovic, G. Salvagiotto *et al.*, 2011 The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity* 34: 175–187.
- ENCODE Project Consortium, 2011 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9: e1001046.
- ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.
- Fan, R., S. Bonde, P. Gao, B. Sotomayor, C. Chen *et al.*, 2012 Dynamic HoxB4-regulatory network during embryonic stem cell differentiation to hematopoietic cells. *Blood* 119: e139–e147.
- Fang, X., J. G. Yoon, L. Li, W. Yu, J. Shao *et al.*, 2011 The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. *BMC Genomics* 12: 11.
- Feng, D., T. Liu, Z. Sun, A. Bugge, S. E. Mullican *et al.*, 2011 A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism. *Science* 331: 1315–1319.
- Fong, A. P., Z. Yao, J. W. Zhong, Y. Cao, W. L. Ruzzo *et al.*, 2012 Genetic and epigenetic determinants of neurogenesis and myogenesis. *Dev. Cell* 22: 721–735.
- Fortschegger, K., P. de Graaf, N. S. Outchkourov, F. M. van Schaik, H. T. Timmers *et al.*, 2010 PHF8 targets histone methylation and RNA polymerase II to activate transcription. *Mol. Cell Biol.* 30: 3286–3298.
- Gao, Z., J. Zhang, R. Bonasio, F. Strino, A. Sawai *et al.*, 2012 PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol. Cell* 45: 344–356.
- Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan *et al.*, 2012 Architecture of the human regulatory network derived from ENCODE data. *Nature* 489: 91–100.
- Gilmour, D. S., and J. T. Lis, 1984 Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl. Acad. Sci. USA* 81: 4275–4279.
- Gilmour, D. S., and J. T. Lis, 1985 In vivo interactions of RNA polymerase II with genes of *Drosophila melanogaster*. *Mol. Cell Biol.* 5: 2009–2018.
- Gotea, V., A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio *et al.*, 2010 Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* 20: 565–577.
- Gowher, H., K. Brick, R. D. Camerini-Otero, and G. Felsenfeld, 2012 Vezf1 protein binding sites genome-wide are associated with pausing of elongating RNA polymerase II. *Proc. Natl. Acad. Sci. USA* 109: 2370–2375.
- Gu, F., H. K. Hsu, P. Y. Hsu, J. Wu, Y. Ma *et al.*, 2010 Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Syst. Biol.* 4: 170.
- Han, J., P. Yuan, H. Yang, J. Zhang, B. S. Soh *et al.*, 2010 Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature* 463: 1096–1100.
- Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew *et al.*, 2011 CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* 43: 630–638.
- He, A., S. W. Kong, Q. Ma, and W. T. Pu, 2011 Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. USA* 108: 5632–5637.
- Hecht, A., S. Strahl-Bolsinger, and M. Grunstein, 1996 Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature* 383: 92–96.
- Heikkinen, S., S. Väisänen, P. Pehkonen, S. Seuter, V. Benes *et al.*, 2011 Nuclear hormone 1 $\alpha$ , 25-dihydroxyvitamin D<sub>3</sub> elicits a genome-wide shift in the locations of VDR chromatin occupancy. *Nucleic Acids Res.* 39: 9181–9193.
- Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin *et al.*, 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38: 576–589.
- Heng, J. C., B. Feng, J. Han, J. Jiang, P. Kraus *et al.*, 2010 The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell* 6: 167–174.
- Ho, J. W., E. Bishop, P. V. Karchenko, N. Nègre, K. P. White *et al.*, 2011 ChIP-chip vs. ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* 12: 134.
- Ho, L., R. Jothi, J. L. Ronan, K. Cui, K. Zhao *et al.*, 2009 An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of



- the core pluripotency transcriptional network. *Proc. Natl. Acad. Sci. USA* 106: 5187–5191.
- Hollenhorst, P. C., K. J. Chandler, R. L. Poulsen, W. E. Johnson, N. A. Speck *et al.*, 2009 DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet.* 5: e1000778.
- Holmstrom, S. R., T. Deering, G. H. Swift, F. J. Poelwijk, D. J. Mangelsdorf *et al.*, 2011 LRH-1 and PTF1-L coregulate an exocrine pancreas-specific transcriptional network for digestive function. *Genes Dev.* 25: 1674–1679.
- Horak, C. E., and M. Snyder, 2002 ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* 350: 469–483.
- Horiuchi, S., A. Onodera, H. Hosokawa, Y. Watanabe, T. Tanaka *et al.*, 2011 Genome-wide analysis reveals unique regulation of transcription of Th2-specific genes by GATA3. *J. Immunol.* 186: 6378–6389.
- Hu, M., J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. S. Qin, 2010 On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.* 38: 2154–2167.
- Hu, G., D. E. Schones, K. Cui, R. Ybarra, D. Northrup *et al.*, 2011 Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res.* 21: 1650–1658.
- Hunkapiller, J., Y. Shen, A. Diaz, G. Cagney, D. McCleary *et al.*, 2012 Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. *PLoS Genet.* 8: e1002576.
- Hutchins, A. P., S. Poulain, and D. Miranda-Saavedra, 2012 Genome-wide analysis of STAT3 binding in vivo predicts effectors of the anti-inflammatory response in macrophages. *Blood* 119: e110–e119.
- Iyer, V. R., C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder *et al.*, 2001 Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409: 533–538.
- Johannes, F., R. Wardenaar, M. Colomé-Tatché, F. Mousson, P. de Graaf *et al.*, 2010 Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* 26: 1000–1006.
- Johnson, D. S., A. Mortazavi, R. M. Myers, and B. Wold, 2007 Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497–1502.
- Joseph, R., Y. L. Orlov, M. Huss, W. Sun, S. L. Kong *et al.*, 2010 Integrative model of genomic factors for determining binding site selection by estrogen receptor- $\alpha$ . *Mol. Syst. Biol.* 6: 456.
- Jung, H., J. Lacombe, E. O. Mazzoni, K. F. Liem, Jr., J. Grinstein *et al.*, 2010 Global control of motor neuron topography mediated by the repressive actions of a single hox gene. *Neuron* 67: 781–796.
- Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando *et al.*, 2010 Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467: 430–435.
- Kassouf, M. T., J. R. Hughes, S. Taylor, S. J. McGowan, S. Soneji *et al.*, 2010 Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res.* 20: 1064–1083.
- Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. *Genome Res.* 12: 996–1006.
- Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park, 2008 Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26: 1351–1359.
- Kim, S. W., S. J. Yoon, E. Chuong, C. Oyulu, A. E. Wills *et al.*, 2011 Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. *Dev. Biol.* 357: 492–504.
- Kim, T. K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear *et al.*, 2010 Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465: 182–187.
- Klisch, T. J., Y. Xi, A. Flora, L. Wang, W. Li *et al.*, 2011 In vivo Atoh1 targetome reveals how a proneural transcription factor regulates cerebellar development. *Proc. Natl. Acad. Sci. USA* 108: 3288–3293.
- Koepfel, M., S. J. van Heeringen, D. Kramer, L. Smeenk, E. Janssen-Megens *et al.*, 2011 Crosstalk between c-Jun and TAp73 $\alpha/\beta$  contributes to the apoptosis-survival balance. *Nucleic Acids Res.* 39: 6069–6085.
- Kong, S. L., G. Li, S. L. Loh, W. K. Sung, and E. T. Liu, 2011 Cellular reprogramming by the conjoint action of ER $\alpha$ , FOXA1, and GATA3 to a ligand-inducible growth state. *Mol. Syst. Biol.* 7: 526.
- Kouwenhoven, E. N., S. J. van Heeringen, J. J. Tena, M. Oti, B. E. Dutilh *et al.*, 2010 Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. *PLoS Genet.* 6: e1001065.
- Krebs, A. R., J. Demmers, K. Karmodiya, N. C. Chang, A. C. Chang *et al.*, 2010 ATAC and Mediator coactivators form a stable complex and regulate a set of non-coding RNA genes. *EMBO Rep.* 11: 541–547.
- Kunars, G., N. Y. Chia, J. Jeyakani, C. Hwang, X. Lu *et al.*, 2010 Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* 42: 631–634.
- Kwon, H., D. Thierry-Mieg, J. Thierry-Mieg, H. P. Kim, J. Oh *et al.*, 2009 Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors. *Immunity* 31: 941–952.
- Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli *et al.*, 2012 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22: 1813–1831.
- Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10: R25.
- Law, M. J., K. M. Lower, H. P. Voon, J. R. Hughes, D. Garrick *et al.*, 2010 ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell* 143: 367–378.
- Lee, B. K., A. A. Bhinge, and V. R. Iyer, 2010 Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res.* 39: 3558–3573.
- Lefterova, M. I., D. J. Steger, D. Zhuo, M. Qatanani, S. E. Mullican *et al.*, 2010 Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Mol. Cell. Biol.* 30: 2078–2089.
- Li, L., R. Jothi, K. Cui, J. Y. Lee, T. Cohen *et al.*, 2010 Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat. Immunol.* 12: 129–136.
- Li, M., Y. He, W. Dubois, X. Wu, J. Shi *et al.*, 2012 Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. *Mol. Cell* 46: 30–42.
- Li, Q., J. Brown, H. Huang, and P. Bickel, 2011 Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* 5: 1752–1779.
- Lieb, J. D., X. Liu, D. Botstein, and P. O. Brown, 2001 Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* 28: 327–334.
- Lin, Y. C., S. Jhunjhunwala, C. Benner, S. Heinz, E. Welinder *et al.*, 2010 A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.* 11: 635–643.
- Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon *et al.*, 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315–322.
- Little, G. H., H. Noshmeh, S. K. Baniwal, B. P. Berman, G. A. Coetzee *et al.*, 2011 Genome-wide Runx2 occupancy in prostate cancer cells suggests a role in regulating secretion. *Nucleic Acids Res.* 40: 3538–3547.
- Liu, W., B. Tanasa, O. V. Tyurina, T. Y. Zhou, R. Gassmann *et al.*, 2010 PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* 466: 508–512.
- Liu, Z., D. R. Scannell, M. B. Eisen, and R. Tjian, 2011 Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* 146: 720–731.
- Lo, K. A., M. K. Baumann, A. P. Baumann, C. J. Donahue, M. A. Thiede *et al.*, 2011 Genome-wide profiling of H3K56 acetylation and transcription factor binding sites in human adipocytes. *PLoS ONE* 6: e19778.
- Lu, F., K. Tsai, H. S. Chen, P. Wikramasinghe, R. V. Davuluri *et al.*, 2012 Identification of host-chromosome binding sites and candidate gene targets for Kaposi's sarcoma-associated herpesvirus LANA. *J. Virol.* 86: 5752–5762.

- Ma, Z., T. Swigut, A. Valouev, A. Rada-Iglesias, J. Wysocka *et al.*, 2010 Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat. Struct. Mol. Biol.* 18: 120–127.
- MacIsaac, K. D., K. A. Lo, W. Gordon, S. Motola, T. Mazor *et al.*, 2010 A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLOS Comput. Biol.* 6: e1000773.
- Mahony, S., E. O. Mazzoni, S. McCuine, R. A. Young, H. Wichterle *et al.*, 2010 Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis. *Genome Biol.* 12: R2.
- Marban, C., T. Su, R. Ferrari, B. Li, D. Vatakis *et al.*, 2011 Genome-wide binding map of the HIV-1 Tat protein to the human genome. *PLoS ONE* 6: e26894.
- Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink *et al.*, 2008 Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134: 521–533.
- Martinez, P., M. Thanasoula, A. R. Carlos, G. Gómez-López, A. M. Tejera *et al.*, 2010 Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites. *Nat. Cell Biol.* 12: 768–780.
- Mazzoni, E. O., S. Mahony, M. Iacovino, C. A. Morrison, G. Mountoufaris *et al.*, 2011 Embryonic stem cell-based mapping of developmental transcriptional programs. *Nat. Methods* 8: 1056–1058.
- McManus, S., A. Ebert, G. Salvagiotto, J. Medvedovic, Q. Sun *et al.*, 2011 The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. *EMBO J.* 30: 2388–2404.
- Mendoza-Parra, M. A., M. Walia, M. Sankar, and H. Gronemeyer, 2011 Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Mol. Syst. Biol.* 7: 538.
- Mendoza-Parra, M. A., M. Van Gool, M. A. Mohamed Saleem, D. G. Ceschin, and H. Gronemeyer, 2013 A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res.* 41: e196.
- Meyer, M. B., P. D. Goetsch, and J. W. Pike, 2012 VDR/RXR and TCF4/ $\beta$ -catenin cisistoms in colonic cells of colorectal tumor origin: impact on c-FOS and c-MYC gene expression. *Mol. Endocrinol.* 26: 37–51.
- Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman *et al.*, 2007 Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553–560.
- Miller, T. W., J. M. Balko, E. M. Fox, Z. Ghazoui, A. Dunbier *et al.*, 2011 ER $\alpha$ -dependent E2F transcription can mediate resistance to estrogen deprivation in human breast cancer. *Cancer Discov.* 1: 338–351.
- Miyazaki, M., R. R. Rivera, K. Miyazaki, Y. C. Lin, Y. Agata *et al.*, 2011 The opposing roles of the transcription factor E2A and its antagonist Id3 that orchestrate and enforce the naive fate of T cells. *Nat. Immunol.* 12: 992–1001.
- Mullen, A. C., D. A. Orlando, J. J. Newman, J. Lovén, R. M. Kumar *et al.*, 2011 Master transcription factors determine cell-type-specific responses to TGF $\beta$  signaling. *Cell* 147: 565–576.
- Mullican, S. E., C. A. Gaddis, T. Alenghat, M. G. Nair, P. R. Giacomini *et al.*, 2011 Histone deacetylase 3 is an epigenomic brake in macrophage alternative activation. *Genes Dev.* 25: 2480–2488.
- Nakayama, S., Y. Kanno, H. Takahashi, D. Jankovic, K. T. Lu *et al.*, 2011 Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* 35: 919–931.
- Nishiyama, A., L. Xin, A. A. Sharov, M. Thomas, G. Mowrer *et al.*, 2009 Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell* 5: 420–433.
- Nitzsche, A., M. Paszkowski-Rogacz, F. Matarese, E. M. Janssen-Megens, N. C. Hubner *et al.*, 2011 RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE* 6: e19470.
- Norton, L., M. Fourcaudot, M. A. Abdul-Ghani, D. Winnier, F. F. Mehta *et al.*, 2011 Chromatin occupancy of transcription factor 7-like 2 (TCF7L2) and its role in hepatic glucose metabolism. *Diabetologia* 54: 3132–3142.
- Novershtern, N., A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining *et al.*, 2011 Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* 144: 296–309.
- Ntziachristos, P., A. Tsigos, P. van Vlierberghe, J. Nedjic, T. Trimarchi *et al.*, 2012 Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat. Med.* 18: 298–301.
- Palii, C. G., C. Perez-Iratxeta, Z. Yao, Y. Cao, F. Dai *et al.*, 2010 Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO J.* 30: 494–509.
- Pehkonen, P., L. Welter-Stahl, J. Diwo, J. Rynänen, A. Wienecke-Baldacchino *et al.*, 2012 Genome-wide landscape of liver X receptor chromatin binding and gene regulation in human macrophages. *BMC Genomics* 13: 50.
- Peng, S., A. A. Alekseyenko, E. Larschan, M. I. Kuroda, and P. J. Park, 2007 Normalization and experimental design for ChIP-chip data. *BMC Bioinformatics* 8: 219.
- Pepke, S., B. Wold, and A. Mortazavi, 2009 Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* 6: S22–S32.
- Ptasinska, A., S. A. Assi, D. Mannari, S. R. James, D. Williamson *et al.*, 2012 Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia* 26: 1829–1841.
- Qi, H. H., M. Sarkissian, G. Q. Hu, Z. Wang, A. Bhattacharjee *et al.*, 2010 Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. *Nature* 466: 503–507.
- Quenneville, S., G. Verde, A. Corsinotti, A. Kapopoulou, J. Jakobsson *et al.*, 2011 In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol. Cell* 44: 361–372.
- Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn *et al.*, 2010 A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470: 279–283.
- Rahl, P. B., C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine *et al.*, 2010 c-Myc regulates transcriptional pause release. *Cell* 141: 432–445.
- Ramagopalan, S. V., A. Heger, A. J. Berlanga, N. J. Mauger, M. R. Lincoln *et al.*, 2010 A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res.* 20: 1352–1360.
- Ramos, Y. F., M. S. Hestand, M. Verlaan, E. Krabbendam, Y. Ariyurek *et al.*, 2010 Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. *Nucleic Acids Res.* 38: 5396–5408.
- Rao, N. A., M. T. McCalman, P. Moulos, K. J. Francoijs, A. Chatziioannou *et al.*, 2011 Coactivation of GR and NF $\kappa$ B alters the repertoire of their binding sites and target genes. *Genome Res.* 21: 1404–1416.
- Remeseiro, S., A. Cuadrado, G. Gómez-López, D. G. Pisano, and A. Losada, 2012 A unique role of cohesin-SA1 in gene regulation and development. *EMBO J.* 31: 2090–2102.
- Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings *et al.*, 2000 Genome-wide location and function of DNA binding proteins. *Science* 290: 2306–2309.
- Rey, G., F. Cesbron, J. Rougemont, H. Reinke, M. Brunner *et al.*, 2011 Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biol.* 9: e1000595.
- Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao *et al.*, 2007 Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4: 651–657.
- Sadasivam, S., S. Duan, and J. A. DeCaprio, 2012 The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes Dev.* 26: 474–489.
- Sahu, B., M. Laakso, K. Ovaska, T. Mirtti, J. Lundin *et al.*, 2011 Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO J.* 30: 3962–3976.
- Sakabe, N. J., I. Aneas, T. Shen, L. Shokri, S. Y. Park *et al.*, 2012 Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum. Mol. Genet.* 21: 2194–2204.
- Schödel, J., C. Bardella, L. K. Sciesielski, J. M. Brown, C. W. Pugh *et al.*, 2012 Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nat. Genet.* 44:420–425, S1–S2.

- Schlesinger, J., M. Schueler, M. Grunert, J. J. Fischer, Q. Zhang *et al.*, 2010 The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet.* 7: e1001313.
- Schmitz, S. U., M. Albert, M. Malatesta, L. Morey, J. V. Johansen *et al.*, 2011 Jarid1b targets genes regulating development and is involved in neural differentiation. *EMBO J.* 30: 4586–4600.
- Schnetz, M. P., L. Handoko, B. Akhtar-Zaidi, C. F. Bartels, C. F. Pereira *et al.*, 2010 CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. *PLoS Genet.* 6: e1001023.
- Sehat, B., A. Tofigh, Y. Lin, E. Trocmé, U. Liljedahl *et al.*, 2010 SUMOylation mediates the nuclear translocation and signaling of the IGF-1 receptor. *Sci. Signal.* 3: ra10.
- Seitz, V., P. Butzhammer, B. Hirsch, J. Hecht, I. Gütgemann *et al.*, 2011 Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma. *PLoS ONE* 6: e26837.
- Shen, T., I. Aneas, N. Sakabe, R. J. Dirschinger, G. Wang *et al.*, 2011 Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. *J. Clin. Invest.* 121: 4640–4654.
- Shukla, S., E. Kavak, M. Gregory, M. Imashimizu, B. Shutinowski *et al.*, 2011 CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479: 74–79.
- Siersbæk, R., R. Nielsen, S. John, M. H. Sung, S. Baek *et al.*, 2011 Extensive chromatin remodelling and establishment of transcription factor hotspots during early adipogenesis. *EMBO J.* 30: 1459–1472.
- Smeenk, L., S. J. van Heeringen, M. Koepel, B. Gilbert, E. Janssen-Megens *et al.*, 2011 Role of p53 serine 46 in p53 target gene regulation. *PLoS ONE* 6: e17574.
- Smith, E. R., C. Lin, A. S. Garrett, J. Thornton, N. Mohaghegh *et al.*, 2011 The little elongation complex regulates small nuclear RNA transcription. *Mol. Cell* 44: 954–965.
- Soccio, R. E., G. Tuteja, L. J. Everett, Z. Li, M. A. Lazar *et al.*, 2011 Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol. Endocrinol.* 25: 694–706.
- Solomon, M. J., P. L. Larsen, and A. Varshavsky, 1988 Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53: 937–947.
- Stadler, M. B., R. Murr, L. Burger, R. Ivanek, F. Lienert *et al.*, 2011 DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480: 490–495.
- Steger, D. J., G. R. Grant, M. Schupp, T. Tomaru, M. I. Lefterova *et al.*, 2010 Propagation of adipogenic signals through an epigenomic transition state. *Genes Dev.* 24: 1035–1044.
- Sun, J., H. Pan, C. Lei, B. Yuan, S. J. Nair *et al.*, 2011 Genetic and genomic analyses of RNA polymerase II-pausing factor in regulation of mammalian transcription and cell growth. *J. Biol. Chem.* 286: 36248–36257.
- Tallack, M. R., T. Whittington, W. S. Yuen, E. N. Wainwright, J. R. Keys *et al.*, 2010 A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.* 20: 1052–1063.
- Tan, P. Y., C. W. Chang, K. R. Chng, K. D. Wansa, W. K. Sung *et al.*, 2011a Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Mol. Cell. Biol.* 32: 399–414.
- Tan, S. K., Z. H. Lin, C. W. Chang, V. Varang, K. R. Chng *et al.*, 2011b AP-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *EMBO J.* 30: 2569–2581.
- Tang, C., X. Shi, W. Wang, D. Zhou, J. Tu *et al.*, 2010 Global analysis of in vivo EGR1-binding sites in erythroleukemia cell using chromatin immunoprecipitation and massively parallel sequencing. *Electrophoresis* 31: 2936–2943.
- Teo, A. K., S. J. Arnold, M. W. Trotter, S. Brown, L. T. Ang *et al.*, 2011 Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.* 25: 238–250.
- Tijssen, M. R., A. Cvejic, A. Joshi, R. L. Hannah, R. Ferreira *et al.*, 2011a Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev. Cell* 20: 597–609.
- Tiwari, V. K., L. Burger, V. Nikolettou, R. Deogracias, S. Thakurela *et al.*, 2011b Target genes of Topoisomerase IIβ regulate neuronal survival and are defined by their chromatin state. *Proc. Natl. Acad. Sci. USA* 109: E934–E943.
- Tiwari, V. K., M. B. Stadler, C. Wirbelauer, R. Paro, D. Schübeler *et al.*, 2011 A chromatin-modifying function of JNK during stem cell differentiation. *Nat. Genet.* 44: 94–100.
- Trompouki, E., T. V. Bowman, L. N. Lawton, Z. P. Fan, D. C. Wu *et al.*, 2011 Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* 147: 577–589.
- Trowbridge, J. J., A. U. Sinha, N. Zhu, M. Li, S. A. Armstrong *et al.*, 2012 Haploinsufficiency of Dnmt1 impairs leukemia stem cell function through derepression of bivalent chromatin domains. *Genes Dev.* 26: 344–349.
- van Heeringen, S. J., W. Akhtar, U. G. Jacobi, R. C. Akkers, Y. Suzuki *et al.*, 2011 Nucleotide composition-linked divergence of vertebrate core promoter architecture. *Genome Res.* 21: 410–421.
- Vermeulen, M., H. C. Eberl, F. Matarese, H. Marks, S. Denissov *et al.*, 2012 Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* 142: 967–980.
- Verzi, M. P., H. Shin, H. H. He, R. Sulahian, C. A. Meyer *et al.*, 2010 Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. *Dev. Cell* 19: 713–726.
- Verzi, M. P., H. Shin, L. L. Ho, X. S. Liu, and R. A. Shivdasani, 2011 Essential and redundant functions of caudal family proteins in activating adult intestinal genes. *Mol. Cell. Biol.* 31: 2026–2039.
- Vilagos, B., M. Hoffmann, A. Souabni, Q. Sun, B. Werner *et al.*, 2012 Essential role of EBF1 in the generation and function of distinct mature B cell types. *J. Exp. Med.* 209: 775–792.
- Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama *et al.*, 2009 ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457: 854–858.
- Vivar, O. I., X. Zhao, E. F. Saunier, C. Griffin, O. S. Mayba *et al.*, 2010 Estrogen receptor beta binds to and regulates three distinct classes of target genes. *J. Biol. Chem.* 285: 22059–22066.
- Wang, D., I. Garcia-Bassets, C. Benner, W. Li, X. Su *et al.*, 2011a Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474: 390–394.
- Wang, H., J. Zou, B. Zhao, E. Johannsen, T. Ashworth *et al.*, 2011b Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. *Proc. Natl. Acad. Sci. USA* 108: 14908–14913.
- Wang, J., J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield *et al.*, 2012 Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22: 1798–1812.
- Wei, L., G. Vahedi, H. W. Sun, W. T. Watford, H. Takatori *et al.*, 2010 Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. *Immunity* 32: 840–851.
- Wei, G., B. J. Abraham, R. Yagi, R. Jothi, K. Cui *et al.*, 2011 Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* 35: 299–311.
- Weinmann, A. S., P. S. Yan, M. J. Oberley, T. H. Huang, and P. J. Farnham, 2002 Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* 16: 235244.
- Welboren, W. J., M. A. van Driel, E. M. Janssen-Megens, S. J. van Heeringen, F. C. Sweep *et al.*, 2009 ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.* 28: 1418–1428.
- Whyte, W. A., S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton *et al.*, 2011 Enhancer decommisioning by LSD1 during embryonic stem cell differentiation. *Nature* 482: 221–225.
- Wilson, N. K., D. Miranda-Saavedra, S. Kinston, N. Bonadies, S. D. Foster *et al.*, 2009 The transcriptional program controlled by the stem cell leukemia gene Scf/Tal1 during early embryonic hematopoietic development. *Blood* 113: 5456–5465.
- Woodfield, G. W., Y. Chen, T. B. Bair, F. E. Domann, and R. J. Weigel, 2010 Identification of primary gene targets of TFAP2C in hormone

- responsive breast carcinoma cells. *Genes Chromosomes Cancer* 49: 948–962.
- Wu, H., A. C. D'Alessio, S. Ito, K. Xia, Z. Wang *et al.*, 2011a Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* 473: 389–393.
- Wu, H., A. C. D'Alessio, S. Ito, Z. Wang, K. Cui *et al.*, 2011b Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev.* 25: 679–684.
- Wu, J. Q., M. Seay, V. P. Schulz, M. Hariharan, D. Tuck *et al.*, 2012 Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS Genet.* 8: e1002565.
- Xiao, S., D. Xie, X. Cao, P. Yu, X. Xing *et al.*, 2012 Comparative epigenomic annotation of regulatory DNA. *Cell* 149: 1381–1392.
- Xu, C., Z. P. Fan, P. Müller, R. Fogley, A. DiBiase *et al.*, 2011 Nanog-like regulates endoderm formation through the Mxtx2-Nodal pathway. *Dev. Cell* 22: 625–638.
- Yaffe, D., and O. Saxel, 1977 Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle. *Nature* 270: 725–727.
- Yang, Y., Y. Lu, A. Espejo, J. Wu, W. Xu *et al.*, 2010 TDRD3 is an effector molecule for arginine-methylated histone marks. *Mol. Cell* 40: 1016–1023.
- Yang, X. P., K. Ghoreschi, S. M. Steward-Tharp, J. Rodriguez-Canales, J. Zhu *et al.*, 2011 Opposing regulation of the locus encoding IL-17 through direct, reciprocal actions of STAT3 and STAT5. *Nat. Immunol.* 12: 247–254.
- Yao, H., K. Brick, Y. Evrard, T. Xiao, and R. D. Camerini-Otero, 2010 Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev.* 24: 2543–2555.
- Yildirim, O., R. Li, J. H. Hung, P. B. Chen, X. Dong *et al.*, 2011 Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* 147: 1498–1510.
- Yoon, S. J., A. E. Wills, E. Chuong, R. Gupta, and J. C. Baker, 2011 HEB and E2A function as SMAD/FOXH1 cofactors. *Genes Dev.* 25: 1654–1661.
- Yu, M., L. Riva, H. Xie, Y. Schindler, T. B. Moran *et al.*, 2009 Insights into GATA-1-mediated gene activation vs. repression via genome-wide chromatin occupancy analysis. *Mol. Cell* 36: 682–695.
- Yu, M., T. Mazor, H. Huang, H. T. Huang, K. L. Kathrein *et al.*, 2012 Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. *Mol. Cell* 45: 330–343.
- Yu, S., K. Cui, R. Jothi, D. M. Zhao, X. Jing *et al.*, 2010 GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells. *Blood* 117: 2166–2178.
- Yuan, P., J. Han, G. Guo, Y. L. Orlov, M. Huss *et al.*, 2009 Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev.* 23: 2507–2520.
- Yun, K., and B. Wold, 1996 Skeletal muscle determination and differentiation: story of a core regulatory network and its context. *Curr. Opin. Cell Biol.* 8: 877–889.
- Zhang, Y., E. V. Laz, and D. J. Waxman, 2011 Dynamic, sex-differential STAT5 and BCL6 binding to sex-biased, growth hormone-regulated genes in adult mouse liver. *Mol. Cell. Biol.* 32: 880–896.
- Zhao, B., J. Zou, H. Wang, E. Johannsen, C. W. Peng *et al.*, 2011a Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. *Proc. Natl. Acad. Sci. USA* 108: 14902–14907.
- Zhao, L., E. A. Glazov, D. R. Pattabiraman, F. Al-Owaidi, P. Zhang *et al.*, 2011b Integrated genome-wide chromatin occupancy and expression analyses identify key myeloid pro-differentiation transcription factors repressed by Myb. *Nucleic Acids Res.* 39: 4664–4679.

*Communicating editor: T. R. Hughes*